Can Translation Memories afford not to use paraphrasing?

Rohit Gupta[1], **Constantin Orăsan**[1], **Marcos Zampieri**[2,3], **Mihaela Vela**[2], **Josef van Genabith**[2,3]
[1]Research Group in Computational Linguistics, University of Wolverhampton, UK
[2]Saarland University, Germany
[3]German Research Center for Artificial Intelligence (DFKI)
{r.gupta, c.orasan}@wlv.ac.uk
{marcos.zampieri, m.vela}@uni-saarland.de
josef.van_genabith@dfki.de

Abstract

This paper investigates to what extent the use of paraphrasing in translation memory (TM) matching and retrieval is useful for human translators. Current translation memories lack semantic knowledge like paraphrasing in matching and retrieval. Due to this, paraphrased segments are often not retrieved. Lack of semantic knowledge also results in inappropriate ranking of the retrieved segments. Gupta and Orăsan (2014) proposed an improved matching algorithm which incorporates paraphrasing. Its automatic evaluation suggested that it could be beneficial to translators. In this paper we perform an extensive human evaluation of the use of paraphrasing in the TM matching and retrieval process. We measure post-editing time, keystrokes, two subjective evaluations, and HTER and HMETEOR to assess the impact on human performance. Our results show that paraphrasing improves TM matching and retrieval, resulting in translation performance increases when translators use paraphrase enhanced TMs.

1 Introduction

One of the core features of a TM system is the retrieval of previously translated similar segments for post-editing in order to avoid translation from scratch when an exact match is not available. However, this retrieval process is still limited to edit-distance based measures operating on surface form (or sometimes stem) matching. Most of the commercial systems use edit distance (Levenshtein, 1966) or some variation of it, e.g. the open-source TM OmegaT[1] uses word-based edit distance with some extra preprocessing. Although these measures provide a strong baseline, they are not sufficient to capture semantic similarity between the segments as judged by humans.

Gupta and Orăsan (2014) proposed an edit distance measure which incorporates paraphrasing in the process. In the present paper, we perform a human-centred evaluation to investigate the use of paraphrasing in translation memory matching and retrieval. We use the same system as Gupta and Orăsan (2014) and investigate the following questions: (1) how much of an improvement can paraphrasing provide in terms of retrieval? (2) What is the quality of the retrieved segments and its impact on the work of human translators? These questions are answered using human centred evaluations.

To the best of our knowledge, this paper presents the first work on assessing the quality of any type of semantically informed TM fuzzy matches based on post-editing time or keystrokes.

2 Related Work

Several researchers have used semantic or syntactic information in TMs, but their evaluations were shallow and most of the time limited to subjective evaluation carried out by the authors. This makes it hard to judge how much a semantically informed TM matching system can benefit a translator.

Existing research (Planas and Furuse, 1999; Hodász and Pohl, 2005; Pekar and Mitkov, 2007; Mitkov, 2008) pointed out the need for similarity

[1]http://www.omegat.org

calculations in TMs beyond surface form comparisons. Both Planas and Furuse (1999) and Hodasz and Pohl (2005) proposed to use lemma and parts of speech along with surface form comparison. Hodasz and Pohl (2005) also extend the matching process to a sentence skeleton where noun phrases are either tagged by a translator or by a heuristic NP aligner developed for English-Hungarian translation. Planas and Furuse (1999) tested a prototype model on 50 sentences from the software domain and 75 sentences from a journal with TM sizes of 7,192 and 31,526 segments respectively. A fuzzy match retrieved was considered usable if less than half of the words required editing to match the input sentence. The authors concluded that the approach gives more usable results compared to Trados Workbench used as a baseline. Hodasz and Pohl (2005) claimed that their approach stores simplified patterns and hence makes it more probable to find a match in the TM. Pekar and Mitkov (2007) presented an approach based on syntactic transformation rules. On evaluation of the prototype model using a query sentence, the authors found that the syntactic rules help in retrieving better segments.

Recently, work by Utiyama et al. (2011) and Gupta and Orăsan (2014) presented approaches which use paraphrasing in TM matching and retrieval. Utiyama et al. (2011) proposed an approach using a finite state transducer. They evaluate the approach with one translator and find that paraphrasing is useful for TM both in terms of precision and recall of the retrieval process. However, their approach limits TM matching to exact matches only. Gupta and Orăsan (2014) also use paraphrasing at the fuzzy match level and they report an improvement in retrieval and quality of retrieved segments. The quality of retrieved segments was evaluated using the machine translation evaluation metric BLEU (Papineni et al., 2002). Simard and Fujita (2012) used different MT evaluation metrics for similarity calculation as well as for testing the quality of retrieval. For most of the metrics, the authors find that, the metric which is used in evaluation gives better score to itself (e.g. BLEU gives highest score to matches retrieved using BLEU as similarity measure).

Keystroke and post-editing time analysis are not new for TM and MT. Keystroke analysis has been used to judge translators' productivity (Langlais and Lapalme, 2002; Whyman and Somers, 1999). Koponen et al. (2012) suggested that post-editing time reflects the cognitive effort in post-editing the MT output. Sousa et al. (2011) evaluated different MT system performances against translating from scratch. Their study also concluded that subjective evaluations of MT system output correlate with the post-editing time needed. Zampieri and Vela (2014) used post-editing time to compare TM and MT translations.

3 Our Approach and Experiments

We have used the approach presented in Gupta and Orăsan (2014) to include paraphrasing in the TM matching and retrieval process. The approach classifies paraphrases into different types for efficient implementation based on the matching of the words between the source and corresponding paraphrase. Using this approach, the fuzzy match score between segments can be calculated in polynomial time despite the inclusion of paraphrases. The method uses dynamic programming along with greedy approximation. The method calculates fuzzy match score as if the appropriate paraphrases are applied. For example, if the translation memory used has a segment "What is the actual aim of this practice ?" and the paraphrase database has paraphrases "the actual" $\Rightarrow$ "the real" and "aim of this" $\Rightarrow$"goal of this", for the input sentence "What is the real goal of this mission ?", the approach will give a 89.89% fuzzy match score (only one word, "practice", needs substitution with "mission") rather than 66.66% using simple word-based edit distance.

In TM, the performance of retrieval can be measured by counting the number of segments or words retrieved. However, NLP techniques are not 100% accurate and most of the time, there is a tradeoff between the precision and recall of this retrieval process. This is also one of the reasons that TM developers shy away from using semantic matching. One cannot measure the gain unless retrieval benefits the translator.

When we use paraphrasing in the matching and retrieval process, the fuzzy match score of a paraphrased segment is increased, which results in the retrieval of more segments at a particular threshold. This increment in retrieval can be classified in two types: without changing the top rank; and by changing the top rank. For example, for a particular input segment, we have two segments A and B in the TM. Using simple edit-distance, A

has a 65% and B has a 60% fuzzy score; the fuzzy score of A is better than that of B. As a result of using paraphrasing we notice two types of score changes:

1. the score of A is still better than or equal to that of B, for example, A has 85% and B has 70% fuzzy score;

2. the score of A is less than that of B, for example, A has 75% and B has 80% fuzzy score.

In the first case, paraphrasing does not supersede the existing model and just facilitates it by improving the fuzzy score so that the top segment ranked using edit distance gets retrieved. However, in the second case paraphrasing changes the ranking and now the top ranked segment is different. In this case, the paraphrasing model supersedes the existing simple edit distance model. This second case also gives a different reference to compare with. We take the top segment retrieved using simple edit distance as a reference against the top segment retrieved using paraphrasing and compare to see which is better for a human translator to work with.

To evaluate the influence of paraphrasing on matching and retrieval, we have carried out four different experiments. Section 3.1 describes the settings and measures used for post-editing evaluation, and Sections 3.2 and 3.3 describe the settings for the subjective evaluations.

3.1 Post-editing Time (PET) and Keystrokes (KS)

In this evaluation, the translators were presented with fuzzy matches and the task was to post-edit the segment in order to obtain a correct translation. The translators were presented with an input English segment, the German segment retrieved from the TM for post-editing and the English segment used for matching in TM.

In this task, we recorded post-editing time (PET) and keystrokes (KS). The post-editing time taken for the whole file is calculated by summing up the time taken on each segment. Only one segment is visible on screen. The segment is only visible after clicking and the time is recorded from when the segment becomes visible until the translator finishes post-editing and goes to the next screen. The next screen is a blank screen so that the translator can have a rest after post-editing a segment. The translators were aware that the time is being recorded. Each translator post-edited half of the segments retrieved using simple edit distance (ED) and half of the segments retrieved using paraphrasing (PP). The ED and PP matches were presented one after the other (ED at odd positions and PP at even positions or vice versa). However, the same translator did not post-edit the match retrieved using PP and ED for the same segment: instead five different translators post-edited the segment retrieved using PP and another five different translators post-edited the match retrieved using ED.

Post-editing time (PET) for each segment is the mean of the normalised time (N) taken by all translators on this segment. Normalisation is applied to account for both slow and fast translators.

$$\text{PET}_j = \frac{\sum_{i=1}^{n} N_{ij}}{n} \tag{1}$$

$$N_{ij} = T_{ij} \times \frac{\text{Avg time on this file by all translators}}{\sum_{j=1}^{m} T_{ij}} \tag{2}$$

In the equations 1 and 2 above, PET_j is the post editing time for each segment j, n is the number of translators, N_{ij} is the normalised time of translator i on segment j, m is the number of segments in the file, and T_{ij} is the actual time taken by a translator i on a segment j.

Along with the post-editing time, we also recorded all printable keystrokes, whitespace and erase keys pressed. For our analysis, we considered average keystrokes pressed by all translators for each segment.

3.2 Subjective Evaluation with Two Options (SE2)

In this evaluation, we carried out subjective evaluation with two options (SE2). We presented fuzzy matches retrieved using both paraphrasing (PP) and simple edit distance (ED) to the translators. The translators were unaware of the details (ED or PP) of how the fuzzy matches were obtained. To neutralise any bias, half of the ED matches were tagged as A and the other half as B, with the same applied to PP matches. The translator has to choose between two options: A is better; or B is better. 17 translators participated in this experiment. Finally, the decision of whether 'ED

is better' or 'PP is better' is made on the basis of how many translators choose one over the other.

3.3 Subjective Evaluation with Three Options (SE3)

This evaluation is similar to Evaluation SE2 except that we provided one more option to translators. Translators can choose among three options: A is better; B is better; or both are equal. 7 translators participated in this experiment.

4 Corpus, Tool and Translators expertise

As a TM and test data, we have used English-German pairs of the Europarl V7.0 (Koehn, 2005) corpus with English as the source language and German as the target language. From this corpus we have filtered out segments of fewer than seven words and greater than 40 words, to create the TM and test datasets. Tokenization of the English data was done using the Berkeley Tokenizer (Petrov et al., 2006). We have used the lexical and phrasal paraphrases from the PPDB corpus (Ganitkevitch et al., 2013) of L size. In these experiments, we have not paraphrased any capitalised words (but we lowercase them for both baseline and paraphrasing similarities calculation). This is to avoid paraphrasing any named entities. Table 1 shows our corpus statistics. The translators involved in

	TM	Test Set
Segments	1565194	9981
Source words	37824634	240916
Target words	36267909	230620

Table 1: Corpus Statistics

our experiments were third year bachelor or masters translation students who were native speakers of German with English language level C1, in the age group of 21 to 40 years with a majority of female students. Our translators were not expert in any specific technical or legal field. For this reason we did not use such a corpus. In this way we avoid any bias from unfamiliarity or familiarity with domain specific terms.

4.1 Familiarisation with the Tool

We used the PET tool (Aziz et al., 2012) for all our human experiments. However, settings were changed depending on the experiment. To familiarise translators with the PET tool we carried out a pilot experiment before the actual experiment with the Europarl corpus. This experiment was done on a corpus (Vela et al., 2007) different from Europarl. 18 segments are used in this experiment. While the findings are not included in this paper, they informed the design of our main experiments.

5 Results and Analysis

The retrieval results are given in Table 2. The table shows the similarity threshold for TM (TH), the total number of segments retrieved using the baseline approach (EDR), the additional number of segments retrieved using the paraphrasing approach (+PPR), the percentage improvement in retrieval obtained over the baseline (Imp), the number of segments that changed their ranking and rose to the top because of paraphrasing (RC), and the number of unique paraphrases used to retrieve +PPR (NP) and RC (NPRC). Table 2 shows that when using

TH	100	[85, 100)	[70, 85)	[55, 70)
EDR	117	98	225	703
+PPR	16	30	98	311
%Imp	13.67	30.61	43.55	44.23
RC	9	14	55	202
NP	24	49	169	535
NPRC	14	24	92	356

Table 2: Results of Retrieval

paraphrasing we obtain around 13.67% increase in retrieval for exact matches and more than 30% and 43% increase in the intervals [85, 100) and [70, 85), respectively. This is a clear indication that paraphrasing significantly improves the retrieval results. We have also observed that there are different paraphrases used to bring about this improvement. In the interval [70, 85), 169 different paraphrases are used to retrieve 98 additional segments.

To check the quality of the retrieved segments human evaluations are carried out. The sets' distribution for human evaluation is given in the Table 3. The sets contain randomly selected segments from the additionally retrieved segments using paraphrasing which changed their top ranking.[2]

TH	100	[85, 100)	[70, 85)	Total
Set1	2	6	6	14
Set2	5	4	7	16
Total	7	10	13	30

Table 3: Test Sets for Human Experiments

[2]The sets are constructed so that a translator can post-edit a file in one sitting. There is no differentiation between the evaluations based on sets and all evaluations are carried out in both sets in a similar fashion with different translators.

	Post-editing				Subjective Evaluations				
	PET		KS		SE2 (2 Options)		SE3 (3 options)		
Seg #	**ED**	**PP**	**ED**	**PP**	**EDB**	**PPB**	**EDB**	**PPB**	**BEQ**
1	42.98	41.30↑	42.4	**0.4**↑	**1**	**16**↑	0	7↑	0
2!+	13.72	10.65↑	2.8	2.4↑	10	7↓	2	2	3
3*!	13.88	12.62↑	2.0	3.6↓	12	5↓	4	1↓	2
4	37.97	**17.64**↑	26.2	**6.2**↑	**1**	**16**↑	0	6↑	1
5!+	21.52	17.69↑	22.4	13.2↑	**13**	**4**↓	2	3↑	2
6!+	41.14	42.74↓	13.2	34.4↓	**4**	**13**↑	2	0	5
7!+	33.69	31.59↑	34.0	33.4↑	10	7↓	1	0	6
8	47.14	**23.41**↑	61.6	**6.4**↑	**0**	**17**↑	0	7↑	0
9	22.89	**14.20**↑	37.2	**2.2**↑	**0**	**17**↑	0	6↑	1
10	46.89	38.20↑	77.6	65.6↑	**1**	**16**↑	0	1	6
11	58.25	53.65↑	82.8	58.8↑	**0**	**17**↑	0	3	4
12!+	34.04	45.03↓	36.8	39.6↓	**2**	**15**↑	0	6↑	1
13	30.34	**21.12**↑	54.8	39.2↑	7	10↑	1	1	5
14!+	75.50	96.54↓	38.8	50.8↓	5	12↑	0	3	4
Set1-subtotal	520.02	466.44	532.60	356.20	66	172	12	46	40
15	24.14	**9.18**↑	24.0	**0.0**↑	5	12↑	1	5↑	1
16*+	28.30	29.20↓	23.4	15.4↑	11	6↓	2	2	3
17*!	65.64	53.49↑	6.2	22.4↓	10	7↓	2	3↑	2
18	41.91	**20.98**↑	28.0	**2.0**↑	**1**	**16**↑	0	6↑	1
19	29.81	19.71↑	23.8	**6.8**↑	7	10↑	2	3↑	2
20	41.25	**15.42**↑	39.0	**3.8**↑	**0**	**17**↑	1	5↑	1
21*!	**42.04**	65.44↓	39.4	36.0↑	7	10↑	1	2	4
22	29.28	35.87↓	17.0	33.4↓	12	5↓	5	0↓	2
23	**32.64**	49.49↓	**11.4**	50.8↓	11	6↓	2	2	3
24!+	59.35	54.54↑	79.6	79.2↑	**17**	**0**↓	5	0↓	2
25	62.51	61.30↑	71.0	54.0↑	**2**	**15**↑	0	3	4
26*!	36.82	41.06↓	55.0	**23.4**↑	**1**	**16**↑	0	6↑	1
27!+	**27.21**	44.02↓	**24.4**	48.8↓	**4**	**13**↑	1	5↑	1
28	40.99	**33.08**↑	39.6	**24.6**↑	5	12↑	3	4↑	0
29	52.01	**31.55**↑	50.6	**23.4**↑	**2**	**15**↑	0	6↑	1
30*!	43.76	38.76↑	38.2	44.6↓	**15**	**2**↓	1	1	5
Set2-subtotal	657.75	603.17	570.6	468.59	110	162	26	53	33
Total	1177.77	1069.61	1103.2	824.79	176	334	38	99	73

Table 4: Results of Human Evaluation on Set1 (1-14) and Set2 (15-30)

Results for human evaluations (PET, KS, SE2 and SE3) on both sets (Set1 and Set2) are given in Table 4. Here 'Seg #' represents the segment number, 'ED' represents the match retrieved using simple edit distance and 'PP' represents the match retrieved after incorporating paraphrasing. 'EDB', 'PPB' and 'BEQ' in Subjective Evaluations represent the number of translators who judge 'ED is better', 'PP is better' and 'Both are equal', respectively.

5.1 Results: Post-editing Time (PET) and Keystrokes (KS)

As we can see in Table 4, improvements were obtained for both sets. ↑ demonstrates cases in which PP performed better than ED and ↓ shows where ED performed better than PP. Entries in bold for PET, KS and SE2 indicate where the results are statistically significant [3].

For Set1, translators made 356.20 keystrokes and 532.60 keystrokes when editing PP and ED matches, respectively. Translators took 466.44 seconds for PP as opposed to 520.02 seconds for ED matches. This means that by using PP matches, translators edit 33.12% less (49.52% more using ED), which saves 10.3% time .

For Set2, translators made 468.59 keystrokes and 570.6 keystrokes when editing PP and ED matches respectively. Translators took 603.17 seconds for PP as opposed to 657.75 seconds for ED matches. This means that by using PP matches, translators edit 17.87% less (21.76% more using ED), which saves 8.29% time.

In total, combining both the sets, translators made 824.79 keystrokes and 1103.2 keystrokes when editing PP and ED matches, respectively. Translators took 1069.61 seconds for PP as opposed to 1177.77 seconds for ED matches. Therefore, by using PP matches, translators edit 25.23% less, which saves time by 9.18%. In other words, ED matches require 33.75% more keystrokes and 10.11% more time. We observe that the percentage improvement obtained by keystroke analysis is smaller compared to the improvement obtained by post-editing time. One of the reasons for this is that the translator spends a fair amount of time reading a segment before starting editing.

[3] $p<0.05$, one tailed Welch's t-test for PET and KS, χ^2 test for SE2. Because of the small sample size for SE3, no significance test was performed on individual segment basis.

5.2 Results: Using post-edited references

We also calculated the human-targeted translation error rate (HTER) (Snover et al., 2006) and human-targeted METEOR (HMETEOR) (Denkowski and Lavie, 2014). HTER and HMETEOR was calculated between ED and PP matches presented for post-editing and references generated by editing the corresponding ED and PP match. Table 5 lists HTER5 and HMETEOR5, which use five corresponding ED or PP references only and HTER10 and HMETEOR10, which use all ten references generated using ED and PP.

Table 5 shows improvements in both the HTER5 and HMETEOR5 scores. For Set-1, HMETEOR5 improved from 59.82 to 81.44 and HTER5 improved from 39.72 to 17.63[4]. For Set-2, HMETEOR5 improved from 69.81 to 80.60 and HTER5 improved from 27.81 to 18.71. We also observe that while ED scores of Set1 and Set2 differ substantially (59.82 vs 69.81 and 39.72 vs 27.81), PP scores are nearly the same (81.44 vs 80.60 and 17.63 vs 18.71). This suggests that paraphrasing not only brings improvement but may also improve consistency.

	Set-1		Set-2	
	ED	**PP**	**ED**	**PP**
HMETEOR5	59.82	81.44	69.81	80.60
HTER5	39.72	17.63	27.81	18.71
HMETEOR10	59.82	81.44	69.81	80.61
HTER10	36.93	18.46	27.26	18.40

Table 5: Results using human targeted references

5.3 Results: Subjective evaluations

The subjective evaluations also show significant improvements.

In subjective evaluation with two options (SE2) as given in Table 4, from a total of 510 (30×17) replies for 30 segments from both sets by 17 translators, 334 replies tagged 'PP is better' and 176 replies tagged 'ED is better' [5].

In subjective evaluation with three options (SE3), from a total of 210 (30×7) replies for 30 segments from both sets by 7 translators, 99 replies tagged 'PP is better', 73 replies tagged 'both are equal' and 38 replies tagged 'ED is better' [6].

[4] For HMETEOR, higher is better and for HTER lower is better

[5] statistically significant, χ^2 test, $p < 0.001$

[6] statistically significant, χ^2 test, $p < 0.001$

5.4 Results: Segment wise analysis

A segment wise analysis of 30 segments from both sets shows that 21 segments extracted using PP were found to be better according to PET evaluation and 20 segments using PP were found to be better according to KS evaluation. In subjective evaluations, 20 segments extracted using PP were found to be better according to SE2 evaluation whereas 27 segments extracted using PP were found to be better or equally good according to SE3 evaluation (15 segments were found to be better and 12 segments were found to be equally good).

We have also observed that not all evaluations correlate with each other on segment-by-segment basis. '!, '+ and '* next to each segment number in Table 4 indicate conflicting evaluations: '!' denotes that PET and SE2 contradict each other, '+' denotes that KS and SE2 contradict each other and '*' denotes that PET and KS contradict each other. In twelve segments where KS evaluation or PET evaluation show PP as statistically significant better, except for two cases all the evaluations also shows them better.[7] For Seg #13 SE3 shows 'Both are equal' and for Seg #26, PET is better for ED, however for these two sentences also all the other evaluations show PP as better.

In three segments (Seg #'s 21, 23, 27) KS evaluation or PET evaluation show ED as statistically significant better, but none of the segment are tagged better by all the evaluations. In Seg #21 all the evaluations with the exception of PET show PP as better. In Seg #23, SE3 shows 'both are equal'. Seg #23 is given as follows:

> **Input:** The next item is the Commission declaration on Belarus .
>
> **ED:** The next item is the Commission Statement on AIDS .//Als nächster Punkt folgt die Erklärung der Kommission zu AIDS.
>
> **PP:** The next item is the Commission statement on Haiti .//Nach der Tagesordnung folgt die Erklärung der Kommission zu Haiti.

In Seg #23, apart from "AIDS" and "Haiti" the source side does not differ but the German side differs. The reason for PP match retrieval was that "statement on" in lower case was paraphrased as "declaration on" while in the other segment "Statement" was capitalised and hence was not paraphrased. If we look at the German side of both ED and PP, "Nach der Tagesordnung" requires a broader context to accept it as a translation of "The next item" whereas "Als nächster Punkt" does not require much context.

In Seg #27, we observe contradictions between post-editing evaluations and subjective evaluations. Seg #27 is given below (EDPE and PPPE are post-edited translations of ED and PP match respectively):

> **Input:** That would be an incredibly important signal for the whole region .
>
> **ED:** That could be an important signal for the future .//Dies könnte ein wichtiges Signal für die Zukunft sein.
>
> **PP:** That really would be extremely important for the whole region .//Und das wäre wirklich für die ganze Region extrem wichtig.
>
> **EDPE:** Dies könnte ein unglaublich wichtiges Signal für die gesamte Region sein.
>
> **PPPE:** Das wäre ein unglaublich wichtiges Signal für die ganze Region.

In subjective evaluations, translators tagged PP as better than ED. But, post-editing suggests that it takes more time and keystrokes to post-edit the PP compare to ED.

There is one segment, Seg #22, on which all the evaluations show that ED is better. Seg #22 is given below:

> **Input:** I would just like to comment on one point.
>
> **ED:** I would just like to emphasise one point.//Ich möchte nur eine Sache betonen.
>
> **PP:** I would just like to concentrate on one issue.//Ich möchte mich nur auf einen Punkt konzentrieren.

In segment 22, the ED match is clearly closer to the input than the PP match. Paraphrasing "on one point" as "on one issue" does not improve the result. Also, "konzentrieren" being a long word takes more time and keystrokes in post-editing.

[7] In this section all evaluations refer to all four evaluations viz PET, KS, SE2 and SE3.

6 Conclusion

Our evaluation answers the two questions previously raised. We conclude that paraphrasing significantly improves retrieval. We observe more than 30% and 43% improvement for the threshold intervals [85, 100) and [70, 85), respectively. The quality of the retrieved segment is also significantly better, which is evident from all our human translation evaluations. On average on both sets used for evaluation, compared to paraphrasing simple edit distance takes 33.75% more keystrokes and 10.11% more time when evaluating the segments who changed their top rank and come up in the threshold intervals because of paraphrasing.

Acknowledgement

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 317471.

References

Aziz, Wilker, S Castilho, and Lucia Specia. 2012. PET: a Tool for Post-editing and Assessing Machine Translation. In *Proceedings of LREC*.

Denkowski, Michael and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of WMT-2014 Workshop*.

Ganitkevitch, Juri, Van Durme Benjamin, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, Atlanta, Georgia.

Gupta, Rohit and Constantin Orăsan. 2014. Incorporating Paraphrasing in Translation Memory Matching and Retrieval. In *Proceedings of EAMT*.

Hodász, Gábor and Gábor Pohl. 2005. MetaMorpho TM: a linguistically enriched translation memory. In *In International Workshop, Modern Approaches in Translation Technologies*.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Koponen, Maarit, Wilker Aziz, Luciana Ramos, and Lucia Specia. 2012. Post-editing time as a measure of cognitive effort. In *Workshop on Post-Editing Technology and Practice in AMTA-2012*, pages 11–20.

Langlais, Philippe and Guy Lapalme. 2002. Trans type: Development-evaluation cycles to boost translator's productivity. *Machine Translation*, 17(2):77–98.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Mitkov, Ruslan. 2008. Improving Third Generation Translation Memory systems through identification of rhetorical predicates. In *Proceedings of LangTech2008*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the ACL*, pages 311–318.

Pekar, Viktor and Ruslan Mitkov. 2007. New Generation Translation Memory: Content-Sensivite Matching. In *Proceedings of the 40th Anniversary Congress of the Swiss Association of Translators, Terminologists and Interpreters*.

Petrov, Slav, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the COLING/ACL*, pages 433–440.

Planas, Emmanuel and Osamu Furuse. 1999. Formalizing Translation Memories. In *Proceedings of the 7th Machine Translation Summit*, pages 331–339.

Simard, Michel and Atsushi Fujita. 2012. A Poor Man s Translation Memory Using Machine Translation Evaluation Metrics . In *Proceedings of AMTA*.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*, pages 223–231.

Sousa, Sheila C.M. de, Wilker Aziz, and Lucia Specia. 2011. Assessing the post-editing effort for automatic and semi-automatic translations of dvd subtitles. In *Proceedings of RANLP*, pages 97–103.

Utiyama, Masao, Graham Neubig, Takashi Onishi, and Eiichiro Sumita. 2011. Searching Translation Memories for Paraphrases. In *Machine Translation Summit XIII*, pages 325–331.

Vela, Mihaela, Stella Neumann, and Silvia Hansen-Schirra. 2007. Querying multi-layer annotation and alignment in translation corpora. In *Proceedings of the Corpus Linguistics Conference CL*.

Whyman, Edward K and Harold L Somers. 1999. Evaluation metrics for a translation memory system. *Software-Practice and Experience*, 29(14):1265–84.

Zampieri, Marcos and Mihaela Vela. 2014. Quantifying the influence of MT output in the translators performance: A case study in technical translation. In *Workshop on Humans and Computer-assisted Translation*.